(12) UK Patent Application (19) GB (11) 2 366 110 (13) A

(43) Date of A Publication 27.02.2002

(21) Application No 0114988.9

(22) Date of Filing 20.06.2001

(30) Priority Data
(31) 09602262 (32) 23.06.2000 (33) US

(71) Applicant(s)
International Business Machines Corporation
(Incorporated in USA - New York)
Armonk, New York 10504, United States of America

(72) Inventor(s)
Paul S Cohen
John R Dildine
Edward J Gleason

(74) Agent and/or Address for Service
J P Richards
IBM United Kingdom Patent Operations, Hursley
Park, Winchester, Hants, SO21 2JN, United Kingdom

(51) INT CL[7]
H04N 7/52

(52) UK CL (Edition T)
H4F FBJ

(56) Documents Cited
EP 0689362 A2 WO99/36918 A1
US 6219640 B US 5880788 A
"A maximum likelihood approach to continuous speech recognition", Bahl et al, IEEE Transactions on Pattern Analysis & Machine Intelligence, Vol, PAMI-5, No 2 (1983).

(58) Field of Search
UK CL (Edition S) H4F FBB FBJ FBK FDX FED
ONLINE DATABASES: WPI, EPODOC, JAPIO, INSPEC.

(54) Abstract Title
Synchronising audio and video.

(57) A method for eliminating synchronisation errors using speech recognition. Using separate audio and visual speech recognition techniques, the method identifies 110 visemes, or visual cues which are indicative of articulatory type, in the video content, and identifies 120 phones and their articulatory types in the audio content. Once the two recognition techniques have been applied, the outputs are compared 130 to determine the relative alignment and, if not aligned, a synchronisation algorithm is applied to time-adjust one or both of the audio and the visual streams in order to achieve synchronisation. Facial features, such as mouth movements, are used to provide visual cues in the video content.

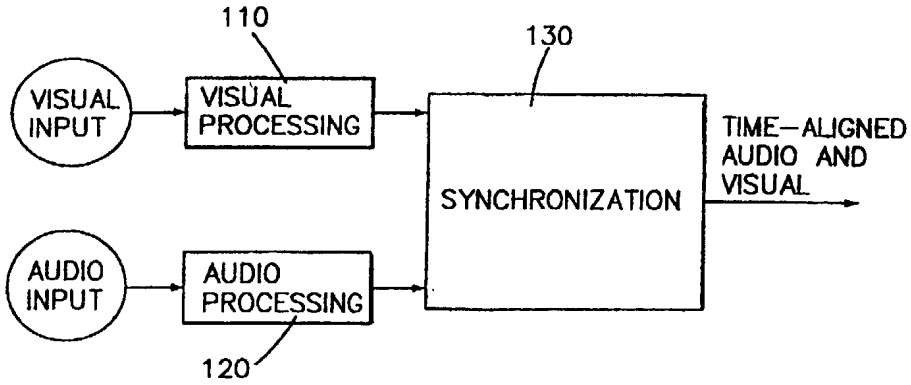

FIG.1

GB 2 366 110 A

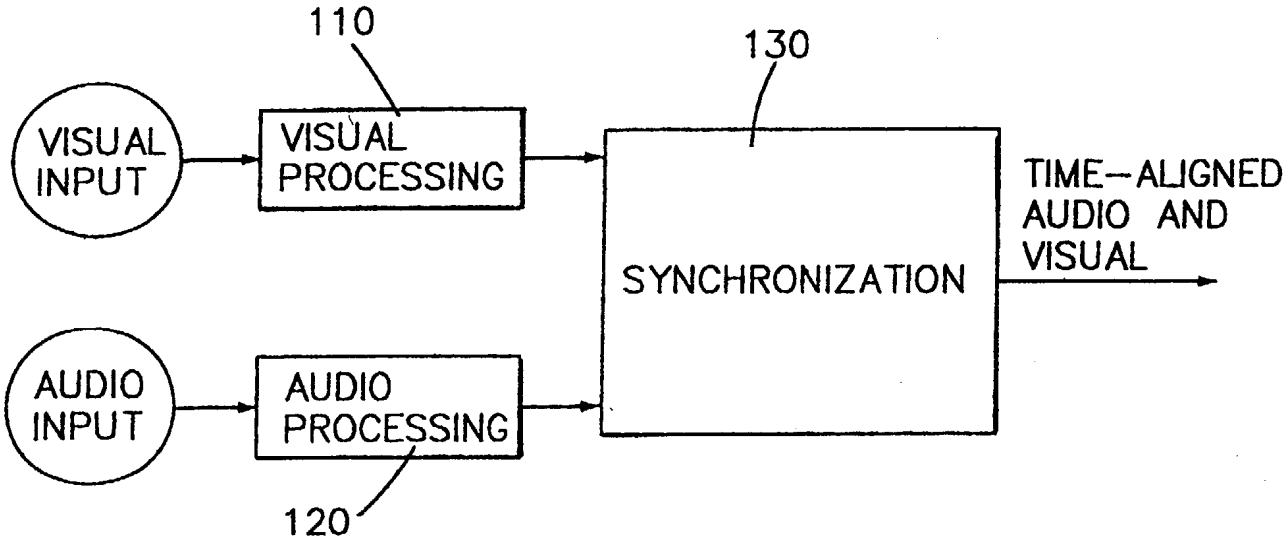
110
130
VISUAL INPUT
VISUAL PROCESSING
SYNCHRONIZATION
TIME-ALIGNED AUDIO AND VISUAL
AUDIO INPUT
AUDIO PROCESSING
120

FIG.1

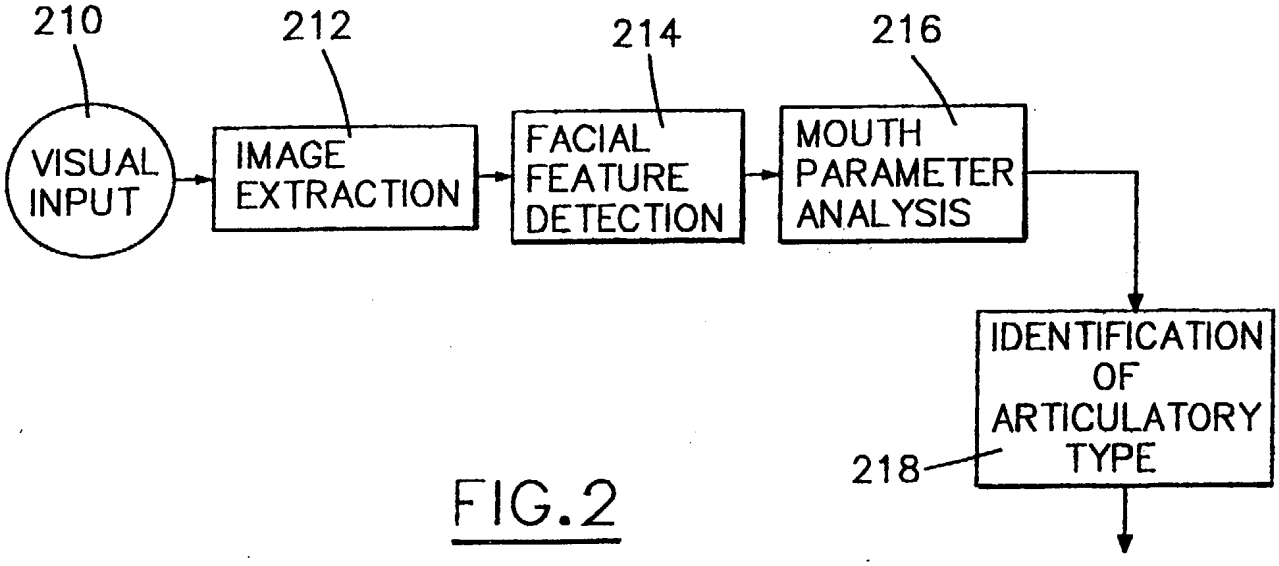
210
212
214
216
VISUAL INPUT
IMAGE EXTRACTION
FACIAL FEATURE DETECTION
MOUTH PARAMETER ANALYSIS
IDENTIFICATION OF ARTICULATORY TYPE
218

FIG.2

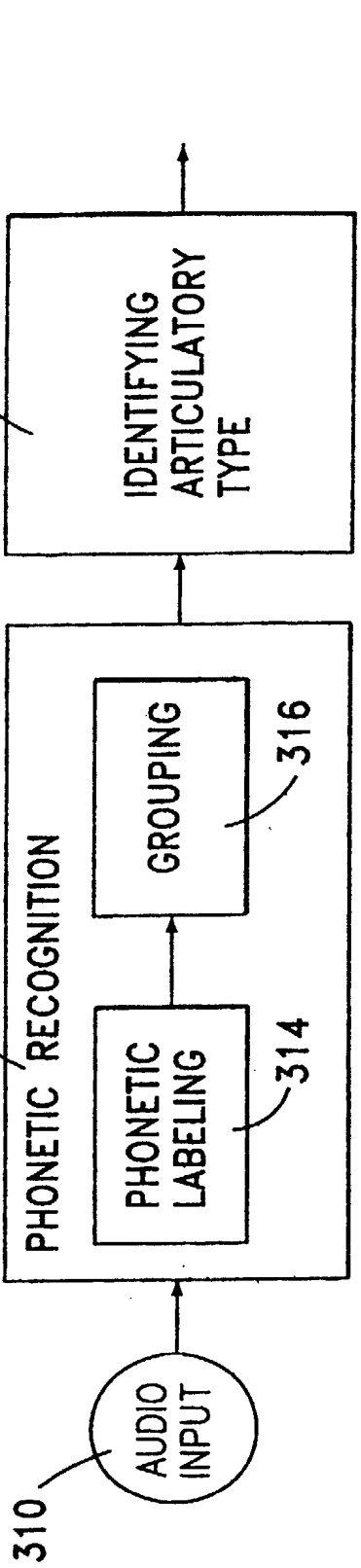
310
AUDIO INPUT
312
PHONETIC RECOGNITION
PHONETIC LABELING
314
GROUPING
316
318
IDENTIFYING ARTICULATORY TYPE

FIG.3

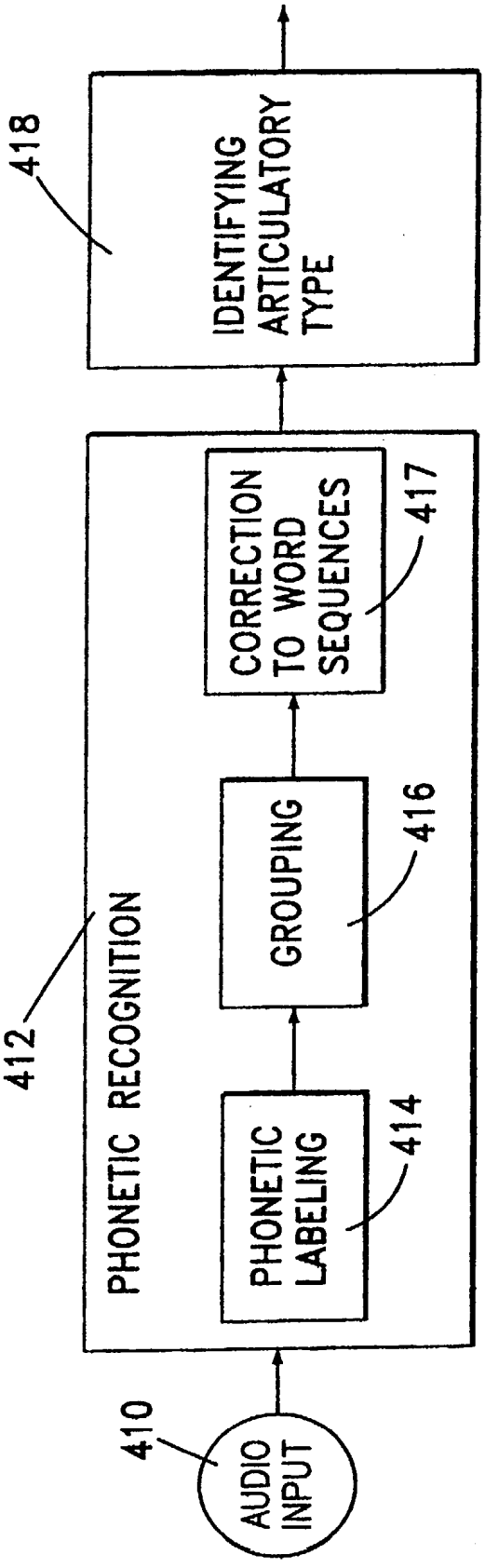
410
AUDIO INPUT
412
PHONETIC RECOGNITION
PHONETIC LABELING
414
GROUPING
416
CORRECTION TO WORD SEQUENCES
417
418
IDENTIFYING ARTICULATORY TYPE

FIG.4

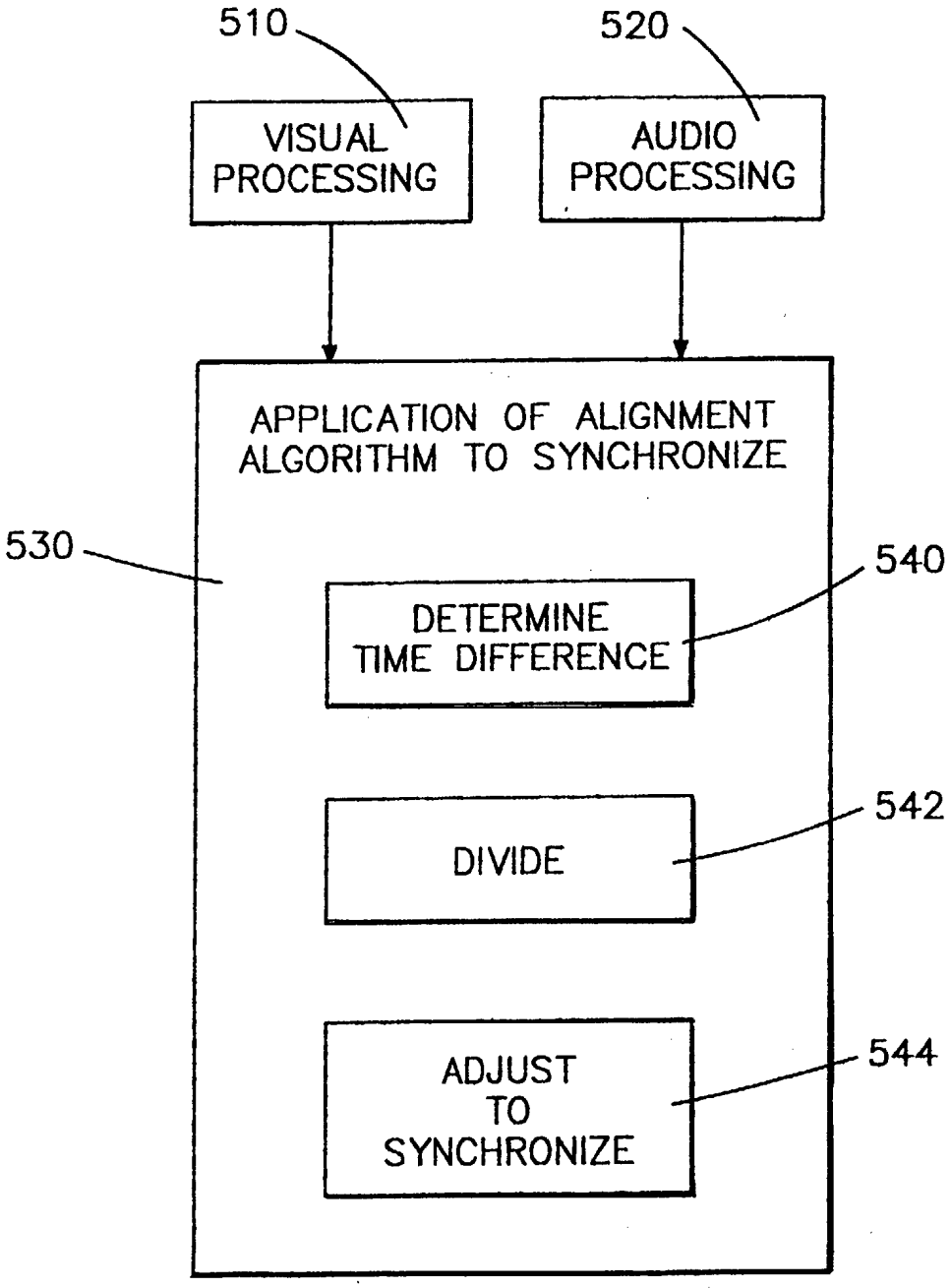
510
520
VISUAL PROCESSING
AUDIO PROCESSING
APPLICATION OF ALIGNMENT ALGORITHM TO SYNCHRONIZE
530
540
DETERMINE TIME DIFFERENCE
542
DIVIDE
544
ADJUST TO SYNCHRONIZE

FIG.5

SYNCHRONISING AUDIO AND VIDEO

The invention relates generally to the synchronisation of audio and video.

A challenge to the effective presentation of multimedia content is time synchronisation, for example the synchronisation of the visual content of a motion picture or video to the corresponding sound. Depending on the exact media involved, the lack of time synchronisation can be caused by a variety of problems. For example, a film loop in a movie theatre or on a television broadcast may have its sound noticeably out of synchronisation with its picture because of operator difficulties in establishing the appropriate tension on the film. Another potential cause of synchronisation errors is a difference in the transmission time between the video, which is often sent via satellite, and the audio, which is often sent over land lines which provide greater security and reliability; which procedure generates an undesirable time differential between picture and audio. Synchronisation problems also commonly occur in live broadcasts when digital effects are inserted into video, causing a time delay in the visual stream but not in the audio stream.

Prior art synchronisation techniques include the use of clapsticks in original double-system filming (i.e., the common practice of recording sound separately from the filmed image stream) to provide a visible and audible cue point to align picture and sound; continuous time code (i.e., a standardised clock reference) recorded with sound and visually recorded on film for time alignment of picture and sound; and, continuous time code integrated into video picture and sound, used to provide time references for synchronisation of picture and sound when they are processed separately, as is commonly the case. Generally, however, prior art synchronisation techniques rely on a human to detect a lack of synchronicity and to attempt to re-align the content.

It is therefore an objective of the present invention to provide an improved system and method for synchronising audio to motion picture or video.

This object is met by the invention claimed in the appended claims.

An embodiment of the invention will now be described, by way of example, with reference to the accompanying drawings, wherein:

Fig. 1 is a schematic representation of a speech recognition system is accordance with an embodiment of the present invention.

Fig. 2 illustrates a representative visual interpretation process flow for use in the present embodiment;

Fig. 3 illustrates a representative audio interpretation process flow for use in the present embodiment;

Fig. 4 illustrates an alternative representative audio interpretation process flow for use in the present embodiment; and

Fig. 5 provides a representative process flow for implementing the present embodiment.

The present embodiment makes use of computer-based audio and visual speech recognition techniques to automatically adjust for synchronisation errors whenever visual facial data and audio/speech data are available in a presentation. The following terms will be used throughout the detailed description.

Allophone- an instantiation of, or a "position variant" of, a phone.

Articulatory type- a speech sound type (e.g., bilabial consonant, labiodental consonant, lip-rounded vowel, open vowel, etc.) which is characterised by specific mouth movements.

Phone- an individual speech sound.

Phoneme- a component of speech comprising multiple allophones, which component functions as a minimum distinctive unit of sound in a linguistic system (e.g., the English language has ~52 phonemes).

Viseme- the minimum distinctive visual manifestation of an acoustic identification (e.g., of an articulatory type). representation in a video or motion picture.

The present embodiment takes advantage of the advancements achieved in the field of visual information, or visemes, in speech recognition, which are the subject of co-pending U.S. Patent application Serial No. 09/452,919 filed December 2, 1999 (IBM Docket No. YO999-428) entitled "Late Integration in Audio-Visual Continuous Speech Recognition" by Verma, et al; patent application Serial No: 09/369,707 (IBM Docket No. YO999-317) entitled "Methods and Apparatus for Audio-Visual Speech Detection and Recognition" by S. Basu, et al; and Serial No: 09/369,706 (IBM Docket No. YO999-318) entitled "Methods and Apparatus for Audio-Visual Speaker Recognition and Utterance Verification" by S. Basu, et al. As detailed therein, visual information, such as the mouth parameters of height, width, and area, along with derivative image information are used to continuously recognise speech, particularly in a non-controlled environment which may have multiple extraneous noise sources. Further to the enhancement of

speech recognition using facial analysis (see: the 09/369,707 application) and the speaker recognition using audio and visual recognition techniques (the 09/369,706 patent application), the Verma patent application focusses on the fusion (or alignment) of data output from a visual recognisor and audio recognisor to improve speech recognition accuracy and to provide automatic speech detection. More particularly, the Verma patent application processes a video signal to identify a class of the most likely visemes found in the signal. Thereafter, the most likely phones and/or phonemes associated with the identified visemes, or with the audio signal, are considered for audio recognition purposes. Therefore, the system and method of the Verma patent applications use both audio and video processing to discern phones produced by the subject, and the phones are, in turn, linked together to discern words.

Under the present invention, audio and video processing are both performed; however, the manner in which the processing proceeds to produce an output, and the use to which the output is made are significantly different from that seen in the Verma patent application. Moreover, the present invention can be implemented in a variety of situations, as will be further detailed below, and not just for continuous recognition of utterances of a live speaker.

Video recognition processing can be conducted in one of a plurality of processing sequences. As described in the aforementioned Verma patent, an image pyramid can first be used for the identification of face candidates, followed by the extraction of a mouth image using facial feature detectors. Once the locating of important facial features has been completed and the mouth image identified, the system performs analysis of the visemes, or visual image components of the mouth image, in order to recognise what speech component is being emitted from the mouth. The aforementioned patent application used probability analysis of the visual information to identify one or more phoneme candidates which are most likely to be associated with the viseme, followed by selection of one of the ~52 English language phonemes. While the Verma patent provides a high degree of speech recognition accuracy, the method is processing-intensive and is not necessarily suitable for use in all instances where time synchronisation is required.

Fig. 1 provides a schematic representation of a speech recognition system in accordance with an embodiment of the present invention. The speech recognition system comprises a visual recognition processing segment

110, an audio recognition processing segment 120, and a synchronisation segment 130 wherein the outputs of components 110 and 120 are time-synchronized (time-aligned). The visual recognition segment 110 comprises components of the type discussed in the aforementioned patent applications, which locate and analyse facial features, and specifically the mouth image, in a visual presentation, as discussed below with reference to Fig. 2, and additionally includes a processing component for identifying the articulatory type of speech which is indicated by the analysis of the mouth image and for providing a video segment output which comprises time-aligned articulatory types.

The audio recognition segment 120 comprises components of the type known in the art, as detailed in the article entitled "A Maximum Likelihood Approach to Continuous Speech Recognition" by L. Bahl, et al, IEEE Transactions on Pattern Analysis and Machine Intelligence (1983). In operation, the audio recognition segment uses a processing component to interpret either intermediate output of a phonetic recognition module or intermediate word output and to generate an audio output which also comprises time-aligned articulatory types. The video segment output and the audio segment output are then provided to the synchronisation segment 130 for appropriate synchronisation based on the articulatory type outputs.

Fig. 2 illustrates a representative visual interpretation process flow for use in the present embodiment. The visual input is introduced at 210 and may be a video signal, a motion picture stream, or a live video stream to which audio must be synchronised. The first step for processing the visual input is the image extraction step which is conducted at 212. As detailed above, the image extraction step may be conducted by pyramid overlay or other equivalent extraction technique. Once the image of interest (i.e., the facial image) has been located at step 212, the image is analysed to detect facial features at 214 and specifically to detect the location of the mouth. Analysis of mouth parameters is conducted at step 216, to determine such relative parameter values as mouth width in relation to height, tongue location, etc. The relative mouth parameter values are sufficient to allow the system to identify which articulatory type is being formed by the mouth at step 218. A database having mouth parameters correlated to articulatory type for the relevant language (e.g., English, French, etc.) is used in step 218 for the identification of articulatory type. The output of the visual speech recognizer is time-stamped with a time stamp so that an identified articulatory type can be readily located for synchronisation.

In accordance with the present embodiment, it is not necessary to identify specific phones or phonemes in order to synchronise the audio to the video as long as the articulatory type of the speech utterance can be identified. To use the English language as an example, it is sufficient for the visual speech recognizer to recognize that, for example, a bilabial consonant (p or b or m), or a labiodental consonant (f or v), or a lip-rounded vowel or semi-vowel (u as in blue, o as in go, w, etc), or an open vowel (short a as in cat, broad a as in father, etc.) is being formed by the mouth. The output of the visual speech recognition component of the present invention therefore, comprises time-stamped articulatory types.

Fig. 3 illustrates one representative audio interpretation process flow for use in the present embodiment. Audio input is provided at 310 for phonetic recognition at step 312. Phonetic recognition comprises identification of phones, which phonetic recognition process is typically implemented by the steps of phonetic labelling at 314 followed by grouping into phones at 316. The phonetic labelling at step 314 comprises labelling each successive time sample, of length typically on the order of 10 msec., as the nearest-matching member of a predetermined set of phonetic prototypes. The grouping step at 316 comprising grouping successive time-sample labels into recognized phones. Each phone, which is time stamped as a result of the processing in 314, is then characterised at step 318 to identify its articulatory type. The output of the Fig. 3 process flow, therefore, comprises time stamped articulatory types which have been identified from the audio input

Fig. 4 illustrates an alternative representative audio interpretation process flow. Audio input is provided at 410 for phonetic recognition at 412. Phonetic labelling is accomplished at 414 and grouping into recognized phones is done at 416. These recognized phones are used in conjunction with a "language model" 417, which is comprised of a grammar or of a table of statistical likelihoods of word sequences or of a combination of these two, to estimate the sequence of words instantiated by the stream of recognized phones. This intermediate output in words (i.e., what is commonly the final output of the audio speech recognizer) is used as the audio starting point for this process. Each word is then looked up in the phonetic-dictionary module (the "baseform lexicon") of the audio processing segment in order to establish a string of time-aligned phones. Then, as in the previous option detailed in Fig. 3, each phone is characterised by a table lookup procedure as to its articulatory type. The output of the Fig.

4 process flow is, therefore, time stamped articulatory types which have been identified from the audio input.

Fig. 5 provides a representative process flow for implementing the present embodiment. The visual processing 510 in visual processing segment (110 of Fig. 1) is conducted on the visual input, independent of, but often simultaneously with, the processing of the audio input at 520 in the audio processing segment (120 of Fig. 1). The time-aligned output of the visual component and the time-aligned output of the audio component are provided for algorithmic alignment at 530 in the synchronisation segment (130 of Fig. 1). The algorithmic alignment, which uses methods which are known in the art such as Viterbi alignment (as discussed in the aforementioned Bahl, et al article, as well as in an article entitled "The Viterbi Algorithm" by G. David Forney, Jr. from the Proceedings of the IEEE (March 1973), performs synchronisation of the time-stamped articulatory type output of the visual speech recognizer with the time-stamped articulatory type output of the audio speech recognizer. The synchronisation can be performed over short or long stretches, as desired, e.g., from a fraction of a second to several hours.

If the audio and video time alignments are sufficiently close in time (e.g., less than or equal to 0.5 of the time taken up by one frame on the audio-visual medium making use of the present embodiment), the audio and visual portions of the presentation are deemed to be synchronised. If not, a synchronisation algorithm, which time adjusts the two portions, is invoked. Preferably, because the change is generally less noticeable and disturbing to the audience, the audio portion is time-adjusted to the visual portion; however if deemed necessary, the visual portion can be time-adjusted to the audio portion; or, each can be time-adjusted at a different rate to achieve alignment. One advantageous version of the synchronisation algorithm, which adjusts the audio and visual portions gradually so as not to be noticeable to the audience operates as follows:

(a) at step 540, determine the difference in time between the leftmost matching articulatory types of the visual and audio speech recognizer output;

(b) at step 542, divide the difference into segments of length equal to 0.5 of the frame length of the relevant audio-visual medium (plus any remainder). {note: the frame length will vary according to the medium. For example, for movies in the U.S., it is 1/24 sec; for video in the U.S., it is approximately 1/30 sec; and for movies and video in Europe, it is 1/25 sec.); and

(c) at step 544, adjust the audio to the visual representation by moving the audio forward or backward (as appropriate) by one segment per frame until all segments (and any remainder) have been moved. Note that the pitch of the speech segments that are moved should be adjusted back to their original frequencies, by means familiar to those skilled in the art, so as to appear natural to the audience.

The present invention has many applications including, but not limited to, the following:

(1) synchronisation of sound and picture on a film loop in movie theatres, on television, etc.;

(2) synchronisation of sound and picture for movies, television, etc. delivered via satellite;

(3) synchronisation of audio to video being transmitted separately from a remote television feed, wherein each station that picks up the feed would have need of the invention as preferably implemented at the individual cable network boxes);

(4) synchronisation of sound and picture for presentations over the internet;

(5) synchronisation of sound and picture for representations in animated productions, such as characters in traditional cartoons and computer-generated characters;

(6) selection among possible synonyms for dubbing of foreign language movies, videotapes, etc. (The synonyms would be coded according to articulatory types in a lexicon. The appropriate word or words would be chosen on the basis of best phonetic match, and then time-aligned according to the methods described herein above);

(7) automation of the laying in of separately recorded sound effects, using nearby speech for time alignment; and

(8) reconstitution of full audio from multiple sound tracks (e.g., removal of undesired timing effects based on distant microphone reception of a performance or news event).

It is noteworthy that the synchronisation algorithm can be applied, as desired, to pre-recorded audio-visual materials; or it can be applied on-the-fly and continuously to, for example, "live" audio-visual materials. Also, although this invention has been described using English-language examples, there is nothing that restricts it to English and it can be implemented for any language. Finally, it should be understood that, while the highest visual recognition accuracy has been realised using facial features linked to speech, it is possible to recognise non-speech acoustic signatures, to link those non-speech acoustic signatures to non-speech

visual "cues" (for example, hand-clapping), to time-stamp the audio and visual output streams, and to synchronise the audio and video based on the identified cues in the time stamped output streams. Under such a visual recognition scenario, the process flow of Fig. 2 would be generalised to the steps of image extraction, feature detection, feature parameter analysis, and correlation of acoustic signatures stored in a database to the feature parameters. For a detailed discussion of the training and use of speech recognition means for identifying audio sources by acoustic signatures, please see co-pending patent application Serial No: 09/602452, (IBM Docket No. YOR9-2000-0130) entitled "System and Method for Control of Lights, Signals, Alarms Using Sound Detection" by W. Ablondi, et al, the teachings of which are herein incorporated by reference.

CLAIMS

1. A method for providing synchronisation of audio to video comprising the steps of:

processing a video signal to generate a video output comprising at least one time stamped acoustic identification of the content of the audio associated with the video signal;

processing an audio signal to generate an audio output comprising at least one time stamped acoustic identification of the content of said audio signal; and

synchronising the video signal to the audio signal by adjusting at least one of the signals to align at least one acoustic identification from the video signal with a corresponding acoustic identification from the audio stream.

2. The method of Claim 1 wherein said synchronising uses a Viterbi algorithm.

3. The method of Claim 1 or 2 wherein said synchronising comprises adjusting the audio signal.

4. The method of Claim 1, 2 or 3 wherein said processing a video signal comprises the steps of:

extracting at least one image from the video signal;

detecting at least one feature in said at least one image;

analysing the parameters of said feature; and

correlating at least one acoustic identification to the parameters of said feature.

5. The method of any preceding claim, wherein each acoustic identification comprises an articulatory type.

6. The method of Claim 4 wherein each acoustic identification comprises an articulatory type and wherein said at least one feature comprises a facial feature.

7. A system for providing synchronisation of audio to video comprising:

a video processing component for processing a video signal to generate a video output comprising at least one time stamped acoustic identification of the content of the audio associated with the video signal;

an audio processing component for processing an audio signal to generate an audio output comprising at least one time stamped acoustic identification of the content of said audio signal; and

a synchronisation component comprising the video signal to the audio signal by adjusting at least one of the signals to align at least one acoustic identification from the video signal with a corresponding acoustic identification from the audio stream.

The Patent Office
PATENTS • DESIGNS • COPYRIGHT • TRADE MARKS
INVESTOR IN PEOPLE

Application No: GB 0114988.9
Claims searched: 1 - 7
Examiner: Matthew Males
Date of search: 10 December 2001

Patents Act 1977
Search Report under Section 17

Databases searched:

UK Patent Office collections, including GB, EP, WO & US patent specifications, in:

UK Cl (Ed.S): H4F FBB, FBJ, FBK, FED, FDX

Int Cl (Ed.7):

Other: Online databases: WPI, EPODOC, JAPIO, INSPEC

Documents considered to be relevant:

Category	Identity of document and relevant passage		Relevant to claims
X, Y	EP 0689362 A2	AT & T - whole document but see abstract.	X: 1, 3 - 7; Y: 2
X	WO99/36918 A1	AVID TECHNOLOGY - whole document but see abstract; page 2, lines 14 - 24.	1, 7 at least
X, P	US 6219640 B1	BASU et al - whole document but see abstract.	1, 7 at least
X	US 5880788	BREGLER - whole document but see abstract; column 2, lines 33 - 48.	1, 7 at least
Y	"A maximum likelihood approach to continuous speech recognition", Bahl et al, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-5, No. 2, (1983): see page 183.		2

X	Document indicating lack of novelty or inventive step	A	Document indicating technological background and/or state of the art.
Y	Document indicating lack of inventive step if combined with one or more other documents of same category.	P	Document published on or after the declared priority date but before the filing date of this invention.
&	Member of the same patent family	E	Patent document published on or after, but with priority date earlier than, the filing date of this application.